FMT_select_validation_predition

Code

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Hide

```
source("pre_processing.R")
```

```
[1] 0.01
[1] 0.01
[1] 0.01
[1] 0.01
```

Hide

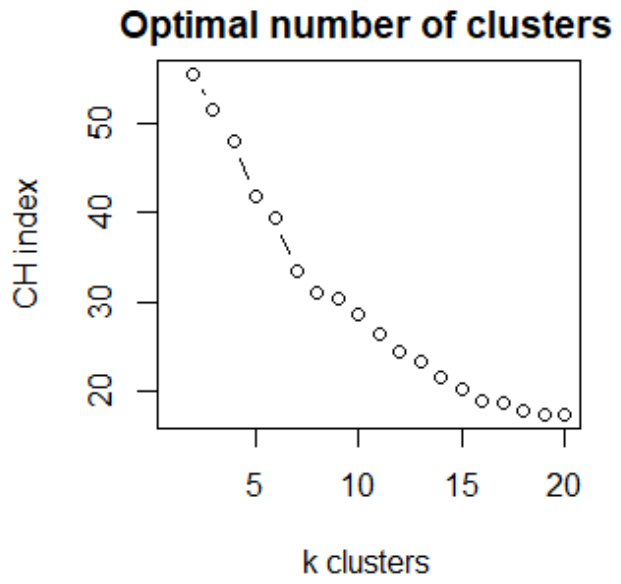
```
##donors
```

Hide

```
###donor entrotpye
don_data <- L6_rela_fil_sAg_others[,unique(c(meta_fil_config$Donor_sra))]/100
# head(colSums(don_data))

don_data_remove = noise.removal(don_data, percent=0.01)
don_data.dist=dist.JSD(don_data_remove)
don_nclusters=NULL
for (k in 1:20) {
  if (k==1) {
    don_nclusters[k]=NA
  } else {
    don_data.cluster_temp=pam.clustering(don_data.dist, k)
    don_nclusters[k]=index.G1(t(don_data_remove), don_data.cluster_temp,  d = don_data.dist,
                         centrotypes = "centroids")
  }
}
```

Hide

```
par(mar= c(4, 5, 2, 2))
layout(matrix(c(1,2), 1, 2, byrow = TRUE), heights = lcm(8))

plot(don_nclusters, type="b", xlab="k clusters", ylab="CH index",main="Optimal number of cluste
rs")
```



Hide

```
don_data.cluster=pam.clustering(don_data.dist, k=2)
# don_nclusters = index.G1(t(don_data_remove), don_data.cluster, d = don_data.dist, centrotypes = "medoids")
don_obs.silhouette=mean(silhouette(don_data.cluster, don_data.dist)[,3])
cat(don_obs.silhouette) #0.1899451
```

```
0.1058641
```

Hide

```
don_obs.pcoa=dudi.pco(don_data.dist, scannf=F, nf=2)
# s.class(don_obs.pcoa$li, fac=as.factor(don_data.cluster), grid=F,sub="Principal coordiante analysis")
```

Hide

```
pdf(file='./figure4/f4_combine_donor_class.pdf')
plot(don_nclusters, type="b", xlab="Number of clusters", ylab="CH index")
dev.off()
```

```
null device
          1
```

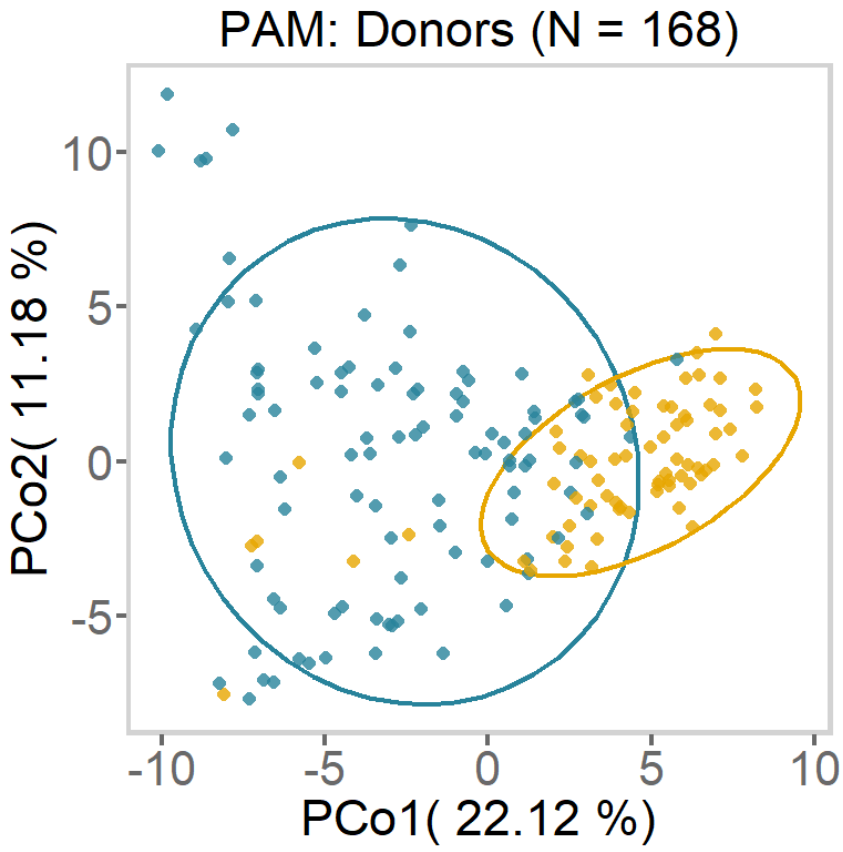
Hide

```
##pcoa

PCo1 <- don_obs.pcoa$li[ ,1]
PCo2 = don_obs.pcoa$li[ ,2]

library(ggplot2)
library(vegan)
# rownames(don_obs.pcoa$li) == don_sra_u[,"SRA"]

Groupn<-'postfmt'

sample.groups <- 1


adonis(don_data.dist ~ don_data.cluster, permutations = 999)
```

```
Call:
adonis(formula = don_data.dist ~ don_data.cluster, permutations = 999)

Permutation: free
Number of permutations: 999

Terms added sequentially (first to last)

                  Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
don_data.cluster   1      2165 2164.99   22.18 0.11787  0.001 ***
Residuals        166     16203   97.61         0.88213
Total            167     18368                 1.00000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hide

```
plotdata <- data.frame(rownames(don_obs.pcoa$li),PCo1,PCo2, sample.groups, don_data.cluster)
colnames(plotdata) <-c("sample","PCo1","PCo2","group", "data.cluster")
pc1 <-floor(don_obs.pcoa$eig[1]*10000/sum(don_obs.pcoa$eig))/100
pc2 <-floor(don_obs.pcoa$eig[2]*10000/sum(don_obs.pcoa$eig))/100
#sample.groups <- don_sra_u[,Groupn]
#shape=factor(substr(don_sra_u[,Groupn], 1, 1))

p<-ggplot(plotdata, alpha=I(0.8))+
  theme_classic()+
  stat_ellipse(aes(x=PCo1, y=PCo2, colour=as.factor(data.cluster),
                   group=as.factor(data.cluster)), level=0.9, size=1.5, show.legend = NA)+
  labs(title=paste("PAM: Donors (N = ", length(PCo1),")", sep=''), x=paste("PCo1(",pc1,"%)"),y=
paste("PCo2(",pc2,"%)") , colour="Cluster")+
  geom_point(aes(x=PCo1,y=PCo2, colour=as.factor(data.cluster), shape=factor(sample.groups),
                 alpha = factor(sample.groups)),size=4)+
  # theme(plot.title = element_text(size=21, hjust = 0.5),
  #     # title=element_text(family ="sans", size=21),
  #       text=element_text(family ="sans", size=18),)+
  theme(text=element_text(family ="sans", size=32), plot.title = element_text(size=34, hjust =
0.5), axis.text = element_text(size=32, color ='dimgray'), axis.title.x = element_text(size=34
), axis.title.y = element_text(size=34), axis.ticks = element_blank())+
  theme(aspect.ratio=0.95, legend.position = c(4, .65), legend.background=element_rect(fill = N
A), legend.text = element_text(size=18))+
  scale_colour_manual(values=c("#28839B", "#E7A600", "#E7A600"))+
  scale_alpha_manual(values = c(0.8))+
  theme(panel.background = element_rect(fill = NA, colour = "lightgrey", size = 3)
        ,axis.line=element_line(colour=NA, size = 0),axis.ticks = element_line(size=1.5, color
 ='dimgray'), axis.ticks.length = unit(7, "pt"));p
```



Hide

```
figli<-1
ggsave(paste("./figure4/4s_donor_entero", ".pdf", sep = ''), device = "pdf")
```

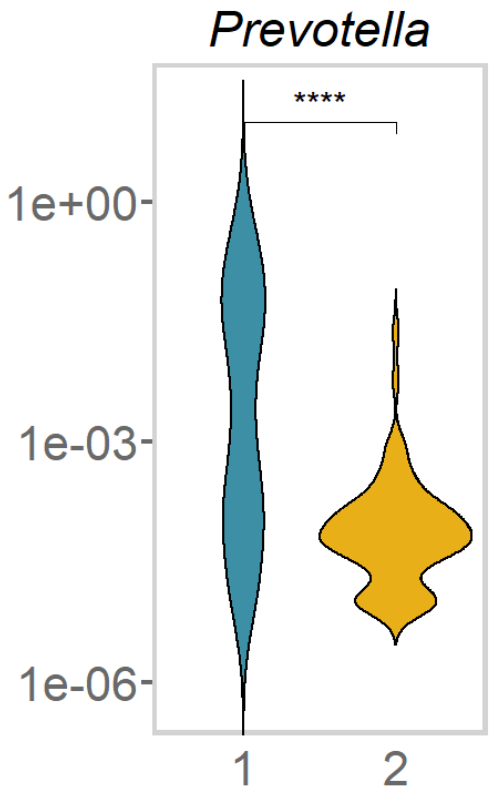
```
Saving 12.9 x 8 in image
```

Hide

```
# figli = figli + 1

# grid.arrange(p1,p2,nrow=1)
beofre_entro_cdi <- cbind(rownames(don_obs.pcoa$li), paste(rep('before', length(don_data.cluste
r)), don_data.cluster, sep = ''), meta_fil_config[match(rownames(don_obs.pcoa$li), meta_fil_con
fig$Previous_sra), 'PRJ'])
```

Hide

```
don_data_mean <- as.data.frame(cbind(sapply(as.character(rownames(don_data_remove)), simp_name
s),
                                     rowMeans(don_data_remove[,don_data.cluster %in% c('1')]),
                                     rowMeans(don_data_remove[,don_data.cluster %in% c('2')])))



don_data_mean[,'V2'] <- as.numeric(as.character(don_data_mean[,'V2']))
don_data_mean[,'V3'] <- as.numeric(as.character(don_data_mean[,'V3']))
# don_data_mean[,'V4'] <- as.numeric(as.character(don_data_mean[,'V4']))


#Prevotella/(Dorea)  Bacteroides

don_simp_name <- sapply(as.character(rownames(don_data_remove)), simp_names)


don_data_remove_c <- rbind(don_data_remove/1000 + 1e-5, don_data.cluster)
rownames(don_data_remove_c) <- c(don_simp_name, 'don_data.cluster')
t_don_data_remove_c<-t(don_data_remove_c)
t_don_data_remove_c <- as.data.frame(t_don_data_remove_c, stringsAsFactors = F)
t_don_data_remove_c$don_data.cluster <- as.character(t_don_data_remove_c$don_data.cluster)


p1<-ggviolin(t_don_data_remove_c, x="don_data.cluster", y="Prevotella", fill = "don_data.cluste
r",#fill = "",
             alpha = 0.9,  add.params = list(alpha=0.3), palette = c("#28839B", "#E7A600"), siz
e = 0.8)+
  stat_compare_means(comparisons = list(c('1', '2')), method = 'wilcox.test', label = "p.signi
f", label.x = 1.5, label.y = 1, size=8)+
  # yscale("log2", .format = FALSE)+
  theme_classic()+theme(legend.position = "right")+xlab(label = '')+ylab(" ")+labs(title='Prevo
tella')+
  theme(text=element_text(family ="sans", size=32), plot.title = element_text(size=34, hjust =
0.5, face = 'italic'), axis.text = element_text(size=32, color ='dimgray'), axis.title.x = elem
ent_text(size=34), axis.title.y = element_text(size=34), axis.ticks = element_blank())+#
  theme(aspect.ratio=2,
        legend.direction = 'horizontal', legend.position = c(5, .15), legend.background=element
_rect(fill = NA)
        ,panel.background = element_rect(fill = NA, colour = "lightgrey", size = 3)
        ,axis.line=element_line(colour=NA, size = 0),axis.ticks.y = element_line(size=1.5, colo
r ='dimgray'), axis.ticks.length = unit(7, "pt"))+
    scale_y_log10(breaks = c(1, 0.001, 0.000001), expand = expansion(add=c(0, 0.2)));p1
```



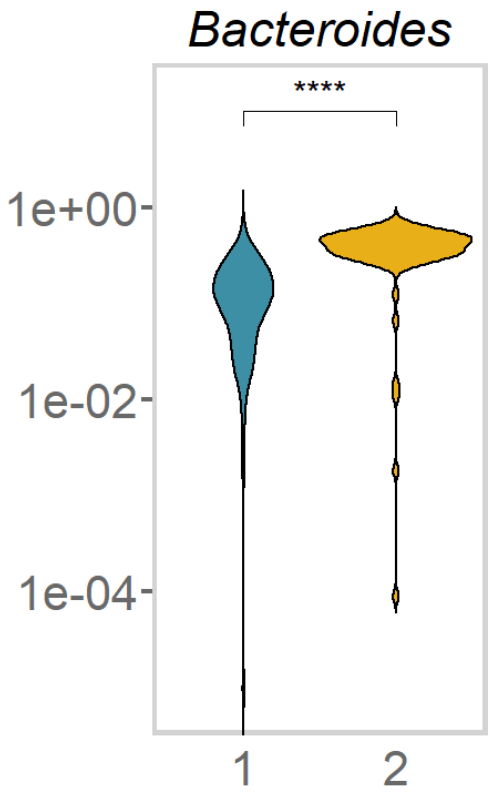
Hide

```
ggsave(paste("./figure4/fig_s4__Donor", 'figli', ".pdf", sep = ''), device = "pdf")
```

```
Saving 12.9 x 8 in image
```

Hide

```
p2<-ggviolin(t_don_data_remove_c, x="don_data.cluster", y="Bacteroides", fill = "don_data.cluster",#fill = "",
             alpha = 0.9,  add.params = list(alpha=0.3), palette = c("#28839B", "#E7A600"), size = 0.8)+
  stat_compare_means(comparisons = list(c('1', '2')), method = 'wilcox.test', label = "p.signif", label.x = 1.5, label.y = 1, size=8)+
  # yscale("log2", .format = FALSE)+
  theme_classic()+theme(legend.position = "right")+xlab(label = '')+ylab(" ")+labs(title='Bacteroides')+
  theme(text=element_text(family ="sans", size=32), plot.title = element_text(size=34, hjust = 0.5, face = 'italic'), axis.text = element_text(size=32, color ='dimgray'), axis.title.x = element_text(size=34), axis.title.y = element_text(size=34), axis.ticks = element_blank())+#, face = 'italic'
  theme(aspect.ratio=2,
        legend.direction = 'horizontal', legend.position = c(5, .15), legend.background=element_rect(fill = NA)
        ,panel.background = element_rect(fill = NA, colour = "lightgrey", size = 3)
        ,axis.line=element_line(colour=NA, size = 0),axis.ticks.y = element_line(size=1.5, color ='dimgray'), axis.ticks.length = unit(7, "pt"))+
    scale_y_log10(expand = expansion(add=c(0, 0.5)));p2
```



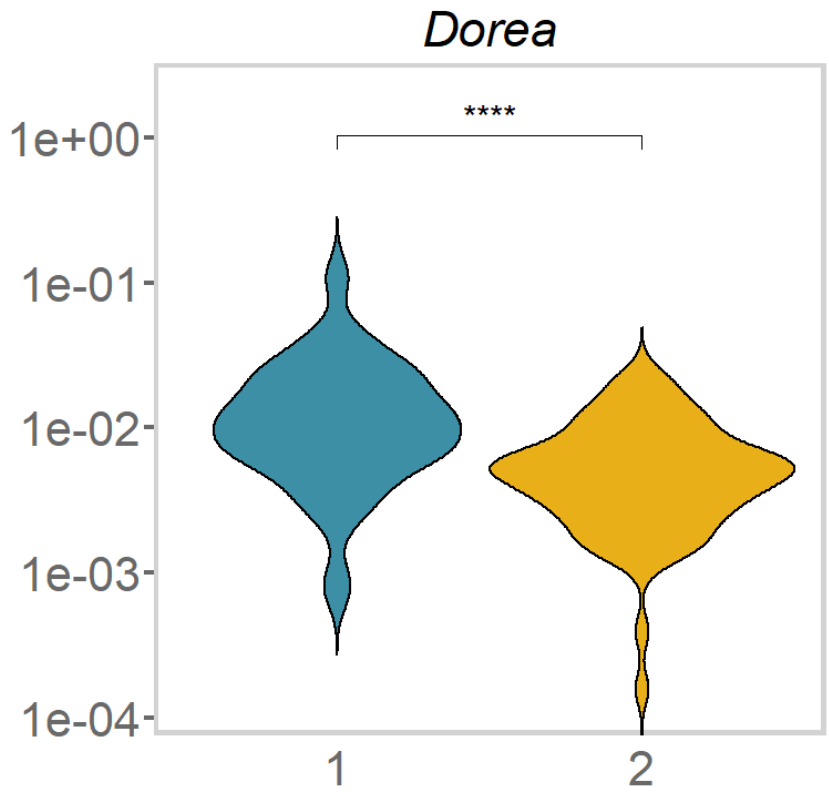
Hide

```
ggsave(paste("./figure4/fig_s4_Donor", 'figli2', ".pdf", sep = ''), device = "pdf")
```

```
Saving 12.9 x 8 in image
```

Hide

```
p3<-ggviolin(t_don_data_remove_c, x="don_data.cluster", y="Dorea", fill = "don_data.cluster",#fill = "",
             alpha = 0.9,  add.params = list(alpha=0.3), palette = c("#28839B", "#E7A600"), size = 0.8)+
  stat_compare_means(comparisons = list(c('1', '2')), method = 'wilcox.test', label = "p.signif", label.x = 1.5, label.y = .01, size=8)+
  # yscale("log2", .format = FALSE)+
  theme_classic()+theme(legend.position = "right")+xlab(label = '')+ylab(" ")+labs(title='Dorea')+
  theme(text=element_text(family ="sans", size=32), plot.title = element_text(size=34, hjust = 0.5, face = 'italic'), axis.text = element_text(size=32, color ='dimgray'), axis.title.x = element_text(size=34), axis.title.y = element_text(size=34), axis.ticks = element_blank())+#, face = 'italic'
  theme(aspect.ratio=1,
        legend.direction = 'horizontal', legend.position = c(5, .15), legend.background=element_rect(fill = NA)
        ,panel.background = element_rect(fill = NA, colour = "lightgrey", size = 3)
        ,axis.line=element_line(colour=NA, size = 0),axis.ticks.y = element_line(size=1.5, color ='dimgray'), axis.ticks.length = unit(7, "pt"))+
    scale_y_log10(expand = expansion(add=c(0, 0.5)));p3
```



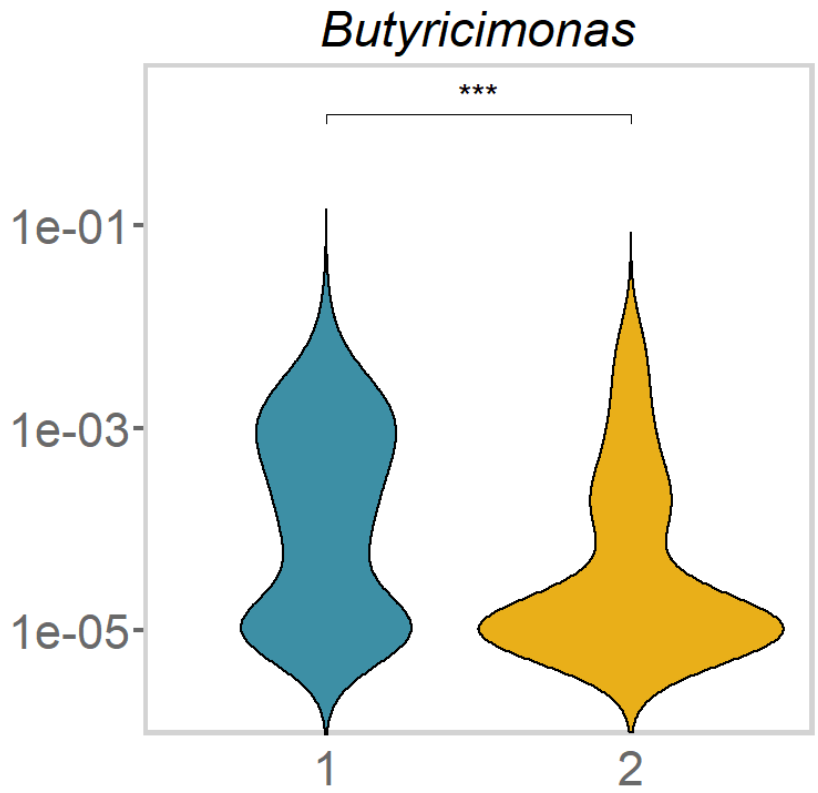
Hide

```
ggsave(paste("./figure4/fig_s4_Donor_dorea", 'figli3', ".pdf", sep = ''), device = "pdf")
```

```
Saving 12.9 x 8 in image
```

Hide

```
ggviolin(t_don_data_remove_c, x="don_data.cluster", y="Butyricimonas", fill = "don_data.cluster",#fill = "",
             alpha = 0.9,  add.params = list(alpha=0.3), palette = c("#28839B", "#E7A600"), size = 0.8)+
  stat_compare_means(comparisons = list(c('1', '2')), method = 'wilcox.test', label = "p.signif", label.x = 1.5, label.y = 1e-1, size=8)+
  # yscale("log2", .format = FALSE)+
  theme_classic()+theme(legend.position = "right")+xlab(label = '')+ylab(" ")+labs(title='Butyricimonas')+
  theme(text=element_text(family ="sans", size=32), plot.title = element_text(size=34, hjust = 0.5, face = 'italic'), axis.text = element_text(size=32, color ='dimgray'), axis.title.x = element_text(size=34), axis.title.y = element_text(size=34), axis.ticks = element_blank())+#, face = 'italic'
  theme(aspect.ratio=1,
        legend.direction = 'horizontal', legend.position = c(5, .15), legend.background=element_rect(fill = NA)
        ,panel.background = element_rect(fill = NA, colour = "lightgrey", size = 3)
        ,axis.line=element_line(colour=NA, size = 0),axis.ticks.y = element_line(size=1.5, color ='dimgray'), axis.ticks.length = unit(7, "pt"))+
    scale_y_log10(expand = expansion(add=c(0, 0.5)))
```



Hide

```
ggsave(paste("./figure4/fig_s4_Donor_buty", 'figlii3', ".pdf", sep = ''), device = "pdf")
```

```
Saving 12.9 x 8 in image
```

Hide

```
source('pre_FMT_sel.R')
```

```
[1] 0.01
[1] 0.01
[1] 0.01
[1] 0.01
[1] 0.05
[1] 0.01
[1] 0.01
[1] 0.01
[1] 0.01
```

Hide

```
repeats <- 501
##navie_run and enterotype_run
##used different features to training in the discovery set
##feature_naive  feature_before  feature_donor  feature_enterotype
```

Hide

```
best_don <- fit_don[[which.min(unlist(lplc_don))]]
ass_don <- apply(mixture(best_don), 1, which.max)
```

Hide

```
feature_before <- feature_naive[startsWith(feature_naive, "B_") | startsWith(feature_naive, "y"
)]
before_result <- navie_run(feature_abun_dat, val_feature_data, feature_before)
```

```
[1] 0.6903636 1.0000000
[1] 0.7079728 1.0000000
[1] 0.6109495 1.0000000
[1] 0.5998929 1.0000000
```

Hide

```
impor_before <- before_result[1]
F1_auc_fil_before<- data.frame(before_result[2])
mean_auc_fil30_before <- data.frame(before_result[3])
mean_val_auc_fil30_before <- data.frame(before_result[4])
mean_predict_fil30_before <- data.frame(before_result[5])
```

Hide

```
feature_donor <- feature_naive[startsWith(feature_naive, "D_") | startsWith(feature_naive, "y"
)]
donor_result <- navie_run(feature_abun_dat, val_feature_data, feature_donor)
```

```
[1] 0.5956166 1.0000000
[1] 0.6047117 1.0000000
[1] 0.5854648 1.0000000
[1] 0.62025 1.00000
```

Hide

```
impor_donor <- donor_result[1]
F1_auc_fil_donor<- data.frame(donor_result[2])
mean_auc_fil30_donor <- data.frame(donor_result[3])
mean_val_auc_fil30_donor <- data.frame(donor_result[4])
mean_predict_fil30_donor <- data.frame(donor_result[5])
```

Hide

```
repeats <- 501
# nfeatures<-20
out_auc_fil30 <- sapply((1:repeats), training_rf, feature_abun_dat=feature_abun_dat, validation
=val_feature_data, feature_b1_d1=c(as.character(mean_impor_b1_d1[1:19, 'id']), 'y'),
feature_b1_d2=c(as.character(mean_impor_b1_d2[1:23, 'id']), 'y'),
feature_b2_d1=c(as.character(mean_impor_b2_d1[1:31, 'id']), 'y'),
feature_b2_d2=c(as.character(mean_impor_b2_d2[1:11, 'id']), 'y'))
#19 23 25 13

out_auc_fil30e <- extract_info(out_auc_fil30, repeats, 10)

auc_fil30 <- data.frame(sapply(1:repeats, function(i){out_auc_fil30e[[i]][1]}))

mean_auc_fil30 <- rowMeans(auc_fil30)
print(mean_auc_fil30[1:2])
```

```
[1] 0.8003142 1.0000000
```

Hide

```
import_b1_d1_fil30 <- data.frame(sapply(1:repeats, function(i){out_auc_fil30e[[i]][3]}))
import_b1_d2_fil30 <- data.frame(sapply(1:repeats, function(i){out_auc_fil30e[[i]][4]}))
import_b2_d1_fil30 <- data.frame(sapply(1:repeats, function(i){out_auc_fil30e[[i]][5]}))
import_b2_d2_fil30 <- data.frame(sapply(1:repeats, function(i){out_auc_fil30e[[i]][6]}))

mean_impor_b1_d1_fil30<-mean_impor(import_b1_d1_fil30)
mean_impor_b1_d2_fil30<-mean_impor(import_b1_d2_fil30)
mean_impor_b2_d1_fil30<-mean_impor(import_b2_d1_fil30)
mean_impor_b2_d2_fil30<-mean_impor(import_b2_d2_fil30)

val_auc_fil30 <- data.frame(sapply(1:repeats, function(i){out_auc_fil30e[[i]][2]}))
mean_val_auc_fil30 <- rowMeans(val_auc_fil30)
print(mean_val_auc_fil30[1:2])
```

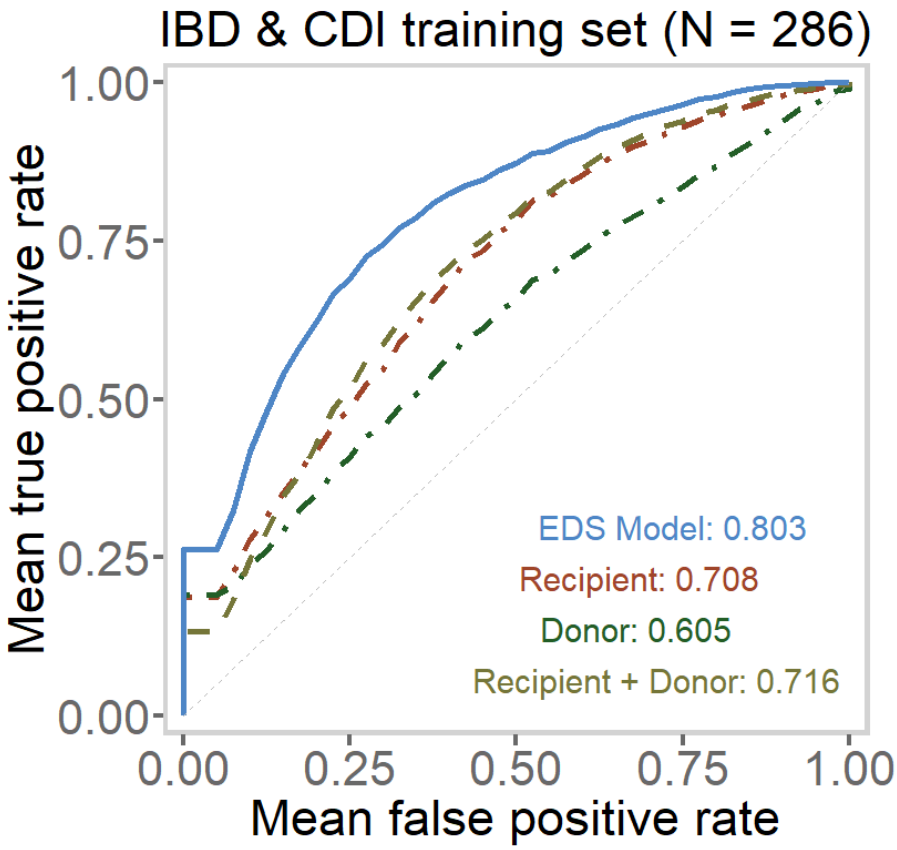
```
[1] 0.7048611 1.0000000
```

Hide

```
F1_fil30 <- data.frame(sapply(1:repeats, function(i){c(out_auc_fil30e[[i]][7])}))

F1_auc_fil <- cbind(t(F1_fil30), t(auc_fil30[1,]), t(val_auc_fil30[1,]))
```

Hide

```
donor_before_after_color <- c("#9F452A", "#4E86C6", "#235E27")
roc_len <- length(seq_roc)
ggplot()+
    scale_x_continuous(expand=c(0,0.03))+
    scale_y_continuous(expand=c(0,0.03))+
    geom_line(aes(seq_roc, c(mean_auc_fil30_before[(3):(2+roc_len),1])), color='#9F452A', size=1.8, linetype=4)+
    geom_line(aes(seq_roc, c(mean_auc_fil30_donor[(3):(2+roc_len),1])), color='#235E27', size=1.8, linetype=4)+
  geom_line(aes(seq_roc, c(mean_auc_fil30_naive[3:(2+roc_len),1])), color='#757639', size=1.8, linetype=2)+
  geom_line(aes(seq_roc, c(mean_auc_fil30[(3):(2+roc_len)])), color='#4E86C6', size=1.8)+
  geom_segment(aes(x=0, xend=1, y=0, yend=1), color='grey', linetype='dashed')+
    geom_text(aes(x=0.685, y=0.4, label=paste('Recipient:', format(round(mean_auc_fil30_before[1,1], 3), nsmall = 3) )), size=8, color='#9F452A', vjust=5)+
    geom_text(aes(x=0.68, y=0.4, label=paste('Donor:', format(round(mean_auc_fil30_donor[1,1], 3), nsmall = 3) )), size=8, color='#235E27', vjust=7)+#087E10
  geom_text(aes(x=0.71, y=0.4, label=paste('Recipient + Donor:', format(round(mean_auc_fil30_naive[1,1], 3), nsmall = 3) )), size=8, color='#757639', vjust=9)+#C6832A
  geom_text(aes(x=0.735, y=0.4, label=paste('EDS Model:', format(round(mean_auc_fil30[1], 3), nsmall = 3) )), size=8, color='#4E86C6', vjust=3)+
  theme_classic()+theme(legend.position = "right")+xlab(label = 'Mean false positive rate')+ylab("Mean true positive rate")+labs(title=paste("IBD & CDI training set ","(N = ",nrow(feature_abun_dat),")", sep = ''))+
   theme(text=element_text(family ="sans", size=32), plot.title = element_text(size=34, hjust = 0.5), axis.text = element_text(size=32, color ='dimgray'), axis.title.x = element_text(size=34), axis.title.y = element_text(size=34),  axis.ticks = element_line(size=1.5, color ='dimgray'), axis.ticks.length = unit(7, "pt"))+
  theme(aspect.ratio = 0.95, legend.background=element_blank()#, legend.position=c(1.75, 0.6)
        ,panel.background = element_rect(fill = NA, colour = "lightgrey", size = 3)
        ,axis.line=element_line(colour="lightgrey")
        ,legend.key = element_rect(fill = NA, color = NA))+
    guides(colour = guide_legend(override.aes = list(size=5)))
```



Hide

```
fig5i = 1
ggsave(paste("./figure5/5main_train", 'fig5i', ".pdf", sep = ''), device = "pdf")
```

```
Saving 12.9 x 8 in image
```

Hide

```
fig5i = fig5i + 1
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.